DOCUMENT RESUME

ED 390 881 TM 023 946

AUTHOR Rudman, Herbert C.; Raudenbush, Stephen W.
TITLE Establishing Optimum Time Limits in the Administration of a Standardized Achievement Test.
PUB DATE 7 Apr 88
NOTE 39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 1988). For related studies, see TM 023 944-945.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Decision Making; Demography; Elementary School Students; Grade 5; Intermediate Grades; Mathematics Achievement; Mathematics Tests; Norms; *Reading Comprehension; Reading Tests; Scores; *Standardized Tests; Student Placement; Teacher Evaluation; Test Construction; Testing Problems; Test Results; *Time; *Timed Tests
IDENTIFIERS Monitoring; *Stanford Achievement Tests

ABSTRACT

A series of three studies was conducted to determine the effects of testing time above or below the recommended time on results of standardized achievement tests with a sample across all three experiments of 1,219 fifth graders in 59 classrooms in Lansing, Michigan. The first two studies considered the effects of increased time; the third explored the point at which the subtest of interest became sensitive to decreased testing time. The first and second studies established that the Reading Comprehension subtest of the Stanford Achievement Test was sensitive to increased time, and that norms lost their utility with increased time. The probability of benefit to teachers because of improved student achievement results and of benefit to students in terms of placement or other instructional decisions was increased. In the third experiment, the Mathematics Applications subtest, which had not been sensitive to increased time, was studied for decreased time, but there were no evident effects. The usual way of establishing the optimum testing time has been based on some predetermined proportion of students who complete the test, often 90%. These studies are a beginning in developing a model to determine the optimum test time more exactly. (Contains 8 tables, 3 figures, and 19 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HERBERT C. RUDMAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ED 390 881

Establishing Optimum Time Limits in the Administration of a Standardized Achievement Test

Herbert C. Rudman and Stephen W. Raudenbush
Michigan State University

BEST COPY AVAILABLE

A paper delivered at the Annual Meeting of the National Council on Measurement in Education, April 7, 1988, New Orleans, LA.

TM023946

Establishing Optimum Time Limits in the Administration of a[1] Standardized Achievement Test

Herbert C. Rudman and Stephen W. Raudenbush
Michigan State University

There is some evidence emerging that the ways used to establish time limits may result in both over-estimates and under-estimates of the time needed to complete various subtests of a standardized achievement test (Rudman & Raudenbush, 1986, 1987, 1988). Methods used to establish maximum time limits for standardized achievement tests employ a survey of times taken by some fixed proportion of students to complete a subtest during an item analysis study (as reported by teachers who administer the test). This technique varies somewhat in that the defined proportion of students may range from 80% to 90%. In some instances, publishers also study the reported proportion of students completing a predetermined percent of items within each subtest. These reported times, then, are pooled across item analysis samples, and a "best estimate" becomes the maximum time to be employed in administering the standardized version of the test.

While this method has proven effective in the past for timed power tests it is not precise enough for contemporary

---

[1] Special appreciation is given to Mr. Sang Jin Kang and Miss Mary Kino for their work in data analysis, and to the administrators, teachers and students of the Lansing, Michigan Public Schools for their cooperation in this investigation during 1986-1988.

use. Resistance to testing has come, in part, from the perceptions of teachers and administrators that too much time is devoted to testing. Two studies recently completed (Rudman & Raudenbush, 1986, 1987) have suggested a research design which may shorten overall achievement testing time while increasing maximum testing time for those subtests found to be time sensitive.

These two studies have employed extended testing time limits to determine the effects of increased time. These investigations have uncovered the possibility that only some subtests remain sensitive to excess testing time. The purpose of a third study recently completed examined the effect upon test scores in a subtest (which previous research has indicated to be insensitive to excess testing time) when the recommended testing time was shortened. This third experiment sought to determine at what point the subtest of interest became sensitive to the effect of decreased testing time.

The first study in this series of investigations (1986) used the Stanford Achievement Test (7th edition) to examine the effects of excess testing time on achievement in Reading Comprehension and Word Study Skills. No significant time effect was noted in the Word Study Skills subtest, but a significant linear effect was noted in Reading Comprehension. These results were discussed in terms of the effect of increased testing time on the use of norms, and on the probabilities that students and teachers would receive

benefits as a result of increased testing time in those subjects that appeared to be time sensitive.

Results indicated that once procedures for establishing norms are violated, the norms lose their practical significance. While no significant change was noted in the norms associated with Word Study Skills, five additional minutes of testing beyond the recommended time could yield, for the Reading Comprehension subtest, an additional .9 grade equivalent score difference and as much as a 1.4 grade equivalent score after fifteen additional minutes. When the same Reading Comprehension subtest was used in a second study (1987), similar results were found.

Concerning the question of the impact that increased testing time could have on the probabilities of "success" of students and teachers in some school-related endeavor, results of the two previous studies in this series consistently indicated that the unit of analysis (data based on individual scores contrasted with group data) is differentially affected by excess time in a subtest which is time sensitive.

Rosenthal and Rubin's (1982) "Binomial effect size display" enables one to translate experimental effect sizes on a continuous variable into probabilities of success on a binary variable. This technique was utilized in the two previous investigations in the series of studies examining testing time determination. Our second study (1987) hypothesized three key decision points; the median, the 80th percentile, and the 20th percentile. Without the benefit of

excess time a randomly selected student would have a probability of .50 of scoring above the median. With excess time that probability would increase to .58 if the test used was time sensitive.

When we hypothesized that a school district wished to place a qualified student into an enrichment program and had set the cut-score at the 80th percentile we estimated that a randomly selected student would have a probability of .20 for acceptance into this program if established time limits were observed. When an additional 15 minutes was used for the time-sensitive Reading Comprehension subtest, the probability of successfully gaining entrance to the enrichment program increased to .31.

If the classroom were the unit of analysis, and if teachers were to be rewarded for classrooms which showed an annual gain that placed them in the top 20% of all classrooms within a district, the probability of success for a randomly selected teacher would rise from .20 to .46 with 15 additional minutes of extra time using a time-sensitive subtest.

We were not, in our earlier studies, nor are we now advocating the violation of procedures established for the administration of standardized achievement tests. We are interested, however, in the consequences of not following standardized recommendations. If testing times are not meticulously followed when scores will be used to make high-stakes decisions, unfair advantages can accrue <u>if</u> the

subtests used are sensitive to additional time. If they are not, then there is no particular advantage to be gained by "fudging" on time limits, nor does there appear to be any great damage either. The point we have made before, and we reiterate again is that the traditional procedures for timing standardized tests may be insufficiently accurate in the context of the high-stakes use increasingly made of them. Since not all subtests appear to be sufficiently desensitized to the effects of time, some tests may well be "under-timed" and others "over-timed", making the strict adherence to time limits essential. On the other hand, we may be unnecessarily saddling test users with longer periods devoted to testing than is necessary.

The Relationship of Test Characteristics to Treatment Effects

The time limits for each of the subtests of the Stanford Achievement Test have been determined by the "90% criterion". Using this standard, testing time is defined as the time elapsed when 90% of students have completed the item analysis edition of the test. Nunally has referred to elapsed time as the "comfortable time limit" (Nunally, 1978, pp. 632-633). This comfortable time includes those answers which have been scored correctly because the examinee either knew the correct response, or else used informed guessing to arrive at the correct response. It is assumed that the remaining 10% of the students would employ random guessing if given more time. Random guessing would not result in significant gains in test scores.

A more rigorous approach to setting time limits for a power test would consider the functional form of the relationship between elapsed time and test scores. If a test is truly a power test, students will make good use of elapsed time up to a point. At that point, the effect of further time will diminish because the students' knowledge relevant to responding to the test items will become exhausted. Thus, the functional form of the relationship between time and test scores ought to be quadratic with negative curvature: the slope of the curve describing that relationship gradually decreases and becomes null. The optimal testing time in a power test is then the point after which further time is unhelpful (See Figure 1).

Insert Figure 1 about here

An important goal of our series of experiments is to discover the optimal testing time by estimating the functional form of the relationship between excess time and test scores for those subtests found susceptible to the effects of excess time. Tests that are assigned time limits less than this optimal time cannot validly be considered power tests. Such tests may be sensitive to variations in test administration procedures and hence be unsuitable for high-stakes use in settings where stringent administrative control is impossible.

Test authors and developers must be concerned with the establishment of optimum time limits so long as tests play a significant role in determining the directions of important policy and instructional decisions. Two dangers exist when testing times are determined for standardized achievement tests. On the one hand, when tests are insufficiently timed, they become vulnerable to variations in test administration. Improper administration of these "time-sensitive" tests can result in inflated norms, erroneous personnel and program decisions, and -- in some instances -- poor allocation of fiscal and human resources. On the other hand, tests which are overtimed can prove to be wasteful of instructional time when less time may yield similar test results.

GENERAL PROCEDURES FOLLOWED IN THE EXPERIMENTAL SERIES

Test Content

The Stanford Achievement Test, (7th edition), Intermediate I, Form F was used to collect student achievement data. Achievement test results for the previous year were used as a covariate. These pretests resulted from the school district's routine administration of Primary 3, Form E battery of the same test series.

The subtests used in Experiment I (1986) were Reading Comprehension, and Word Study Skills. A third summary score of these two subtests was also used (Total Reading). Experiment II (1987) again used the Reading Comprehension subtest and introduced the Mathematic Applications subtest as a new element to be investigated. Experiment III reexamined

the Mathematics Applications subtests. In sum, then, we collected two years of data on the Reading Comprehension subtest, two years of data on Mathematics Applications, and one year's data on Word Study Skills.

Sample and Design

Lansing, Michigan is an urban school district which consists of 33 elementary schools. In each of the three studies, faculty members of these elementary schools volunteered their fifth grade classrooms and students to serve as participants. A total of 59 classrooms supplied usable data for 1,219 pupils across all three experiments. Table 1 describes the samples on key demographic variables and test scores.

Insert Table 1 about here

Each of the classrooms was assigned to one of several blocks that approximated some measure of socio-economic status. Four treatment groups representing testing-time allotments were established. Within each of the blocks, classrooms were assigned at random to one of the four treatments. In each experiment an attempt was made to have a balanced randomized block design (Kirk, 1982, Chapter 6), but extenuating circumstances only permitted a close approximation to such a design.

The blocks were developed to be as homogeneous as possible on prior achievement test scores and on

socioeconomic status as indicated by the proportion of families receiving aid to dependent children. The blocking also ensured that no two classrooms within the same school would experience the same treatment. The blocking variable was viewed as an ordinal variable, ranking the classrooms on background variables related to the outcomes. Table 2 represents the testing-time used within the four treatment groups.

Insert Table 2 about here

Treatment Group 1 used the normal maximum time limit stipulated in the test administration manuals which accompany the Stanford Achievement Test. Treatment Groups 2, 3, and 4 were incremented in 5 minute intervals for the total testing-time allocated across the subtest in Experiments I and II, and decremented in Experiment III. Unadjusted and adjusted posttest means and sample sizes are provided in Tables 4 -6. The maximum time limit for each of the subtests used served as Treatment 1., and all other treatments were either incremented or decremented by 5 minute intervals for a total of 15 minutes. The authors of the Stanford Achievement Test caution that a maximum time is never to be extended (Gardner, Rudman, Karlsen, and Merwin, 1982) for regular testing procedures.

ANALYSIS AND RESULTS OF THREE EXPERIMENTS

Method of Statistical Analysis

In each of the three experiments, the basic design involved pupils nested within block-by-treatment combinations. To increase the statistical power of the analysis, covariates were used. Thus, the analysis employed was a hierarchical (nested) randomized blocks analysis of covariance. Since the data were mildly unbalanced and included classrooms as a random factor, an appropriate method of analysis was a sequential "fitting of constants" analysis. This analysis involves fitting a series of regression models of increasing complexity, each time computing the residual sums of squares. The simplest model fits just a grand mean and enables estimation of the total variation in the outcome. The second simplest model fits the covariate(s), enabling an estimation of the variation left unexplained by the covariate. The effects of blocks, the linear trend, the quadratic trend, and the cubic trend are then sequentially added, and, in each case, the residual sum of squares computed. Finally, a model is fit which includes the covariate(s) and the classroom effects.

This series of models enables a partitioning of the covariate adjusted variation within and between classrooms. The between classroom variation is then further partitioned into variation explained by the blocks and the polynomial trend components. The algorithm for fitting the models and computing the ANOVA table is displayed in Table 3, which

shows also how the degrees of freedom were partitioned using Experiment I as an example. In each of the experiments the residual between classrooms mean square (labelled "blocks by treatments" since each classroom is defined by a block-by-treatment cross-classification) supplies the appropriate error term for the F-test.

Insert Table 3 about here

After completion of the analysis of data from each experiment taken separately, it was deemed useful to conduct a pooled analysis of the two-years data for which Reading Comprehension had been used as an outcome. For this combined analysis, we utilized the hierarchical linear model program developed by Bryk, Raudenbush, Congdon, and Seltzer (1986) which is explained in detail in Raudenbush and Bryk (1986). This method enables maximum likelihood of variance and covariance components in hierarchical designs. One of the key advantages is that a unique solution to the estimation of treatment effects results. The fitting of constants method described above can produce results which may vary depending on the order in which effects are included in the model and the residual variation computed.

Results for Experiment I

Recall that in the first experiment, two outcome variables were employed: Reading Comprehension and Word Study skills. There were four treatment groups defined by having either 0, 5, 10, or 15 extra minutes to complete the test.

The key finding was a highly significant positive, linear effect of excess time on the Reading Comprehension test scores, $F(1,11) = 23.33$, $p < .01$. There was no significant linear effect for Word Study Skills. In fact, there were no significant differences among the treatment means for that outcome $F(3,11) = 2.92$. These results may be found in Table 4. The partial correlation associated with the linear trend was $r = .17$ for the unadjusted scores; while for the covariate adjusted scores, $r = .26$. Considering the variation among class means as the criterion, the partial correlation was $r = .67$ for the adjusted means. These effects were of sufficient magnitude to conclude that excess time could substantively effect decisions about students, and especially, decisions about teacher effectiveness (see Discussion).

Insert Table 4 about here

Largely because of the linear effect of excess time on Reading Comprehension, a significant, positive, linear effect of excess time was also manifest for Total Reading, $F(1,11) = 12.98$, $p < .01$. These results are reported in Table 4, which shows the analysis of variance table and the adjusted treatment means.

Perhaps of equal significance to the linear trend overall was the lack of any quadratic trend. The appearance of a quadratic trend would have enabled estimation of an

optimal timing for the test by providing evidence about when the effect of excess time disappears. The fact that no such quadratic effect appeared indicated that students were, on average, able to continue to make productive use of excess time even after 15 extra minutes spent on the test! This result suggested that if one wishes to view the Reading Test as a pure power test, it is seriously undertimed.

An anomalous result was the appearance of a significant cubic effect of excess time for Reading Comprehension, $F(1,11) = 9.75$, $p < .01$, and for Total Reading, $F(1,11) = 10.86$, $p < 01$. There was no apparent explanation for this result and it served as a source of further motivation for a replication study.

Results of Experiment II

Recall that Experiment II also involved the implementation of four treatments using the same intervals of excess time as in Experiment I. An important difference was that this time the outcome variables were Reading Comprehension and Math Applications. We reasoned that the Math Applications subtest has a structure somewhat analogous to Reading Comprehension (requiring the reading of a passage and then the answering of questions) and that an effect of excess time was plausible.

The key finding in this experiment was the replication of the significant positive linear trend in the case of the Reading Comprehension subtest, $F(1,18) = 8.09$, $p < .02$. There was no significant linear effect for Mathematics

Applications, and in fact, there were no significant differences among treatment means on that subtest, $F(3,18) = .62$. The analysis of variance table and the adjusted means are presented in Table 5.

Insert Table 5 about here

For the Reading Comprehension subtest, the partial correlation between time and the unadjusted scores was $r = .11$; between time and the adjusted scores, $r = .17$. Considering the adjusted classroom means as the criterion, the partial correlation was $r = .52$. Again, these results may be viewed as substantively significant for the effects of excess time on decisions. Once again, perhaps surprisingly, no evidence of a quadratic effect emerged, reaffirming the notion that even at 15 minutes of excess time some students were productively using their time.

Results of Experiment III

Our motivation for conducting the latest experiment was to discover whether a test found insensitive to the effects of excess time might be overtimed. If so, it would also be interesting to discover an "optimal time," that is, a testing time which would be adequate for most students (in the sense that they could make little or no use of more time) but not excessive (in the sense that unproductive testing time would be minimized. For this purpose we chose the Math Applications subtest. No effect of excess time was manifest for this subtest so there was no reason to believe it was

undertimed. This time the four treatment groups had 0, 5, 10, or 15 minutes **less** time than conventionally allocated to complete the test. We predicted a positive linear effect of testing time as well as a quadratic effect with negative (decelerating) curvature. We hoped that the shape of the curve would help us discover an optimal testing time.

Surprisingly, no significant effect of testing time appeared. In fact there were no significant differences among treatment means, $F(3,10) = .39$ (Table 6). Indeed, in comparing results on Math Applications from Experiment II and Experiment III there was no evidence to infer any effect of time whatsoever. This seemed highly surprising, given that the testing time available ranged over a 30 minute interval across the two years.

Insert Table 6 about here

Combined Analysis Involving Reading Comprehension

In both Experiment I and Experiment II, we had found a significant linear effect of excess time for the Reading Comprehension subtest. In Experiment I there had been a seemingly anomalous cubic effect; this effect was absent for the Experiment II results.

To assess the generalizability of the effect of excess time, we conducted a pooled analysis. For both years we used the Total Reading pretest as a covariate. We sought to estimate the combined linear, quadratic and cubic effects;

the year effect (the effect of the year in which the study was conducted), and the year by time interaction effect. The year-by-time interaction effect is highly relevant to the issue of generalizability. A significant year-by-time interaction would indicate a result which is for some reason particular to one replication of the experiment. On the other hand, a significant main effect of time without an interaction would indicate a more stable result.

The hierarchical linear model program enables the analyst to test the homogeneity of regression of the covariate. If heterogeneity is found, it may then be modeled explicitly. No evidence of heterogeneity of regression was found. The estimated variance of the regression coefficients (Table 7) was .002, and the chi-square test with 41 degrees of freedom was 54.98 , $p = .07$. As a result the effect of the covariate was treated as fixed.

Table 7 provides the year-by-time results. There was no significant main effect of year, $t(34) = -.525$. There was a highly significant main effect for the linear trend, $t(34) = 5.18$, $p < .0001$ and no significant linear-by-year interaction effect, $t(34) = -1.32$. For the quadratic effect there was no evidence of either a main effect, $t(34) = -.281$, or of a year-by-quadratic interaction effect, $t(34) = -.06$. In the case of the cubic trend, there was a marginally significant main effect, $t(34) = 1.98$, p .05 and a marginally significant interaction, $t(34) = -2.00$, $p = .05$.

The key results of the pooled analysis were two. First, the linear effect of time was replicated. Both experiments may be viewed as estimating the same linear effect of excess time. Hence the pooled regression coefficient, b = .723 should be viewed as the reliable effect in both experiments instead of the effects estimated in the separate studies. Second, the cubic effect was best viewed as ephemeral; there was no evidence of its replication. That is, given the lack of a significant effect at year 2, the significant year-by-cubic interaction may be viewed as evidence that the effect did not hold up over replication.

Insert Table 7 about here

DISCUSSION

Our initial interest in the effect of test-taking time on test scores arose from the high-stakes use currently being made of standardized achievement tests (Florida Statute, 1985; Lewis, 1985; Tirozzi, 1985; Turlington, 1985). Achievement test scores have increasingly been coupled to educational policies and instructional decisions that seriously affect students and their teachers. Since tests are playing a key role in policy development they have, naturally, been subject to greater scrutiny by those affected. If tests are used to drive curriculum and personnel decisions, and they are, all aspects of test

construction need to be reexamined to ensure that the data yielded by these tests are indeed appropriate for the uses to which they are put.

A key testing procedure is the development of maximum time-limits established for each of the subtests used. Little has been written, in the past twenty years, of investigations into the relationship between the establishment of time limits on standardized achievement tests and test scores. The last intensive review of this problem was a symposium published in Educational and Psychological Measurement, (1960), 20, 221-274. Test developers have generally accepted the method of establishing time limits based on observations of some predetermined proportion of students who complete the test. Although we have referred to the convention identified as the "90% criterion", this proportion and method has varied somewhat among test authors and publishers. The procedure has generally been satisfactory for the more traditional purposes to which standardized achievement tests have been put; analyzing curricular strengths and weaknesses, monitoring individual and group achievement, reporting to parents, and the like. All of these purposes can be thought of as low-risk uses, with little serious consequences.

Standardized achievement tests are being used today in ways that strain the ethical limits of acceptable test use as well as the psychometric limits of the tests (Rudman, 1985a; Rudman, 1985b; Hoover, 1984; AERA-APA-NCME, 1984; Traub,

1983). Some school districts have reported using gain scores derived from repeated use of these tests to award school personnel cash bonuses, (Florida Statute, 1985) and others are placing considerable emphasis upon teacher evaluation by using the results of standardized achievement tests. Given this prevailing picture of test use, cautions have been raised about maintaining a strict adherence to maximum time limits to avoid any possibility of undue advantage being gained when these time limits are not adhered to. If faculty and other staff personnel are to be rewarded in some tangible fashion based upon annual increases in the test scores of students, then any factor in the administration of these tests which can influence a test score and, hence, the probability of reward becomes a target for close scrutiny.

A few points of interest stand out in this series of experimental studies. While we have, to this point in time, only examined three subject areas, we are not convinced that the prevailing method of determining time limits for standardized tests is precise enough for the types of decisions that are influenced by the scores these tests yield. Secondly, while the strong linear effect we have observed in the Reading Comprehension subtest may be an artifact of a particular standardized test battery, we cannot assume that there are no time-sensitive subtests contained within a standardized achievement test. Our work has led us to the development of a model by which we can judge that point at which a test becomes more or less sensitive to time.

While our last study should have shown some sensitivity to time no significant difference was found. We are puzzled by this, but feel that further analysis with other samples is called for. In a series of replication studies chance factors can influence the experimental outcomes; the cubic effect we found in Reading Comprehension in Experiment I but did not find in Experiment II serves as an example of this likelihood.

While our work is not finished we would hope that others will join in exploring the relationship between time and test score. Work is presently underway with other subtests and other standardized achievement tests. We have raised more questions than we have answered, but hopefully these past three years have offered a design for further study, data for further analysis, as well as the introduction of some statistical tools with which to examine more precise ways of establishing time limits for standardized tests.

REFERENCES

AERA-APA-NCME (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., & Congdon, R.T. (1986). An Introduction to HLM: Computer Program and User's Guide. Chicago: University of Chicago, Department of Education.

District Quality Instruction Incentive Program. Sec. 231.532 (3) (F), Florida Statute, (1985).

Gardner, E.F., Rudman, H.C., Karlsen, B., & Merwin, J.C. (1982). Stanford Achievement Test. Primary 3, Form E and Intermediate 1, Form F. Cleveland, OH: The Psychological Corporation.

Hoover, H.D. (1984). The most appropriate scores for measuring educational development in the elementary schools: ge's. Educational Measurement: Issues and Practices, 3, 8-14.

Kirk, R.E. (1982). Experimental Design: Procedures for the Behavioral sciences. Belmont, CA: Brook-Cole Publishers.

Lewis, A.C. (1985). Test scores, test scores on the wall ... Phi Delta Kappan, 66, 387-388.

Morrison, E.J. (1960). On test variance and the dimension of the measurement situation. Educational and Psychological Measurement. 20, 231-250.

Nunnaly, J. (1978). Psychometric Theory. New York: McGraw-Hill Book Company.

Raudenbush, S.W. & Bryk, A.S. (1986). A hierarchical model for studying school effects. Sociology of Education, 59, 1-17.

Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

Rudman, H.C. (1985). Responsible and ethical test use: it cuts both ways. Paper presented at the annual meeting of the National Council on Measurement, and the American Educational Research Association. (a).

Rudman, H.C. (1985). Testing beyond minimums. Occasional Paper No. 5. Springfield, IL: Association of State Assessment Programs. (b).

Rudman, H.C. & Raudenbush, S.W. (1986). The effect of exceeding prescribed time limits in the administration of standardized achievement tests. Research Series No.5. East Lansing, MI: Department of Counseling, Educational Psychology, and Special Education, Michigan State University.

Rudman, H.C. & Raudenbush, S.W. (1987). The effect of exceeding prescribed time limits in the administration of a standardized test of reading comprehension and mathematics applications. Research Series No.6. East Lansing, MI: Department of Counseling, Educational Psychology, and Special Education, Michigan State University.

Rudman, H.C. & Raudenbush, S.W. (1988). Establishing optimum time limits in the administration of a standardized achievement test. A paper presented to the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Tirozzi, G.N. et al. (1985). How testing is changing education in Connecticut. Educational Measurement: Issues and Practice, 4, 12-16.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (ed). Applications of Item Response Theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Turlington, R.E. (1985). How testing is changing education in Florida. Educational Measurement in Education: Issues and Practice, 4, 9-11.

Table 1
Description of the Samples Used in
Experiments I, II, and III (1986, 1987, 1988)

Demographic Variables

Variable	Level	Relative Frequency		
		1986	1987	1988
Sex	Male	45.7	45.5	46.9
	Female	54.3	54.5	53.1
Ethnicity	White (non-Hispanic)	59.9	60.1	56.4
	Other	40.1	39.9	43.6
Family Configuration	One-parent	35.0	38.1	37.7
	Two-parent	65.0	61.9	62.3
Parent Education	Elementary	NA	2.0	3.3
	Junior High	NA	3.7	4.0
	High School	NA	16.3	13.2
	H.S. Graduate	NA	40.3	35.9
	College Attendance	NA	24.5	21.6
	College Graduate	NA	7.9	13.9
	Post-Graduate	NA	5.2	8.1
Native Language	English	62.6	90.5	90.7
	Other	16.4	9.5	9.3
	Missing	21.0	----	----

Test Data

Subtest	Pretest						Postest					
	I		II		III		I		II		III	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
Reading Comp.	36.70	13.80	35.40	12.57	---	---	41.80	11.71	41.42	10.19	---	---
Math Appl.	---	---	25.80	8.58	27.01	8.47	---	---	28.07	7.58	27.05	8.41

Table 2
Testing Time by Treatment and Subtest

Treatment	Reading Comprehension	Math Appl. (I)	Math Appl. (II)
1	30	35	35
2	35	40	30
3	40	45	25
4	45	50	20

Table 3
Analytic Method

Model/Source	df residual	Sum of square residual
Constant	408	(1)
Constant, covariate	407	(2)
Constant, covariate, blocks	403	(3)
Constant, covariate, blocks, linear trend	402	(4)
Constant, covariate, blocks, linear, quadratic trends	401	(5)
Constant, covariate,blocks linear, quadratic, cubic trends	400	(6)
Constant, covariate, classrooms	389	(7)

Partitioning of variation

	df reduction	Sum of squa reductio
Covariate	1	(1) -
Between classes (adjusted)	18	(2) -
Blocks	4	(2) -
Treatments	3	(3) -
Linear	1	(3) -
Quadratic	1	(4) -
Cubic	1	(5) -
Residual (blocks by treatments)	11	(6) -
Within classes (adjusted)	389	(7)
Total	408	(1)

Table 4
ANCOVA Source Tables for (a) Reading comprehension,
(b) Word Study Skills, and (c) Total Reading

Source	df	Sum of Squares	Mean Square	F	Eta.	Eta.
Reading Comprehension						
Covariate	1	33010.977	33010.977		.77	---
Between classes	18	3559.434	197.746		.25	.39
Blocks	4	529.954	132.448		.10	.15
Treatments	3	2273.811	757.937	11.03**	.20	.31
Linear	1	1602.809	1602.809	23.33**	.17	.26
Quadratic	1	1.025	1.025	.01	.00	.01
Cubic	1	669.977	669.977	9.75**	.11	.17
Residual	11	755.669	68.697	1.38*	.12	.18
Within classes	389	19367.745	49.79			
Total	408	55938.156	137.103			
Word Study Skills						
Covariate	1	29658.803	29658.803		.81	---
Between classes	18	1828.834	101.602		.20	.34
Blocks	4	205.673	51.418		.07	.11
Treatments	3	713.729	239.910	2.92	.13	.21
Linear	1	41.564	41.564	.51	.03	.05
Quadratic	1	166.765	166.765	.03	.26	.10
Cubic	1	511.400	511.400		.11	.18
Residual	11	903.432	82.130	2.36**	.14	.24
Within classes	389	13755.967	34.825			
Total	408	45243.604	109.187			

BEST COPY AVAILABLE

Table 5
ANCOVA Source Tables for (a) Reading comprehension, and
(b) Mathematics Applications

Source	df	Sum of Squares	Mean Square	F	Eta$_a$	Eta$_b$
Reading Comprehension						
Covariate	1	27211.561	2741.560		.77	---
Between classes	22	2259.332	102.697		.25	.39
Blocks	1	330.972	330.972		.10	.15
Treatments	3	607.018	202.369	2.76*	.11	.17
Linear	1	594.006	639.003	8.09**	.11	.17
Quadratic	1	12.983	12.983	.18	.00	.01
Cubic	1	.122	.122	.00		
Residual	18	1321.252	73.403	1.70**		
Within classes	447	19315.711	43.212			
Total	470	48786.603	103.801			
Mathematics Applications						
Covariate	1	16477.230	16477.230			
Between classes	22	802.767	36.489			
Blocks	1	1.919	1.919			
Treatments	3	75.224	25.075	.62		
Linear	1	22.750	22.750	.56		
Quadratic	1	0.117	0.117	.00		
Cubic	1	52.356	52.356	1.30		
Residual	18	725.623	40.312	1.85*		
Within classes	447	9733.819	21.776			
Total	470	27013.826	57.476			

* p<.10
** p<.05
*** p<.02
a based on raw posttest score
b based on covariance adusted posttest score

ANCOVA source tables (continued)

Total Reading

Covariate	1	112433.764	112433.764		.82	---
Between Classes	18	8696.347	483.130		.23	.39
Blocks	4	394.744	98.68		.05	.08
Treatments	3	5701.979	1900.660	8.04**	.18	.32
Linear	1	3067.600	3067.600	12.98**	.13	.23
Quadratic	1	68.567	68.567	.29	.02	.03
Cubic	1	2565.812	2565.812	10.86**	.12	.21
Residual	11	2599.624	236.329	1.94**	.12	.21
Within classes	389	47342.032	121.702			
Total	408	168472.083	406.937			

* .10<p<.25
** p<.05
*** p<.01
a based on raw posttest score
b based on covariance adjusted posttest score

Adjusted Treatment Means (Unstandardized)

Subtest	Excess Time			
	None	5 minutes	10 minutes	15 minutes
Reading Comprehension	38.89	42.80	41.09	44.91
Word Study Skills	42.74	43.95	41.20	44.65
Total Reading	80.31	86.89	82.52	90.01

Adjusted Treatment Means (Standardized)

Reading Comprehension	-.29	.08	-.06	.27
Word Study Skills	-.04	.08	-.18	.14
Total Reading	-.23	.10	-.12	.25

Table 5 (continued)

Subtest	Excess Time			
	None	5 minutes	10 minutes	15 minutes
Adjusted Treatment Means (Unstandardized)				
Reading Comprehension	39.53	41.44	42.14	43.03
Mathematics Appl.	27.67	28.52	27.79	28.49
Adjusted Treatment Means (Standardized)				
Reading Comprehension	-.186	.002	.070	.158
Mathematics Appl.	-.053	.059	-.038	.056

Table 6

ANCOVA Source Tables for Mathematics Applications

Source	df	Sum of Squares	Mean Square	F
Covariate	2	16554.814	8277.41	398.91***
Between Class	16	745.84	46.62	2.25**
Block	3	218.5524	72.85	1.54
Treatments	3	55.2847	18.42	.39
Linear	1	7.8232	7.82	.16
Quadratic	1	19.354	19.35	.41
Cubic	1	28.1075	28.11	.60
Residual	10	472.1029	47.21	2.28**
Within Classes	321	6662.27	20.75	
Total	339	23863.04705	70.69	

** P<.01
*** P.001

Table 7: Results of pooled Analysis (Experiments I and II) for Reading Comprehension

Effect	Coefficient	Standard Error	t (df = 34)
	(a) Fixed Effects		
Covariate	.390	.011	34.65
Linear	.723	.140	5.18
Quadratic	-.089	.317	-.28
Cubic	.283	.143	1.98
Year	-.166	.317	-.52
Linear x Year	-.184	.140	-1.32
Quad. x Year	-.020	.317	-----
Cub. x Year	-.287	.143	-2.01

(b) Random Effects

Effect	Variance	Chi-square	df	P
Adjusted Class Means	1.88	282.31	34	000

Table 8
A Comparison of Mean Raw Scores and Derived Scaled Scores
and Grade Equivalents by Treatment Level[1]
and Experiment

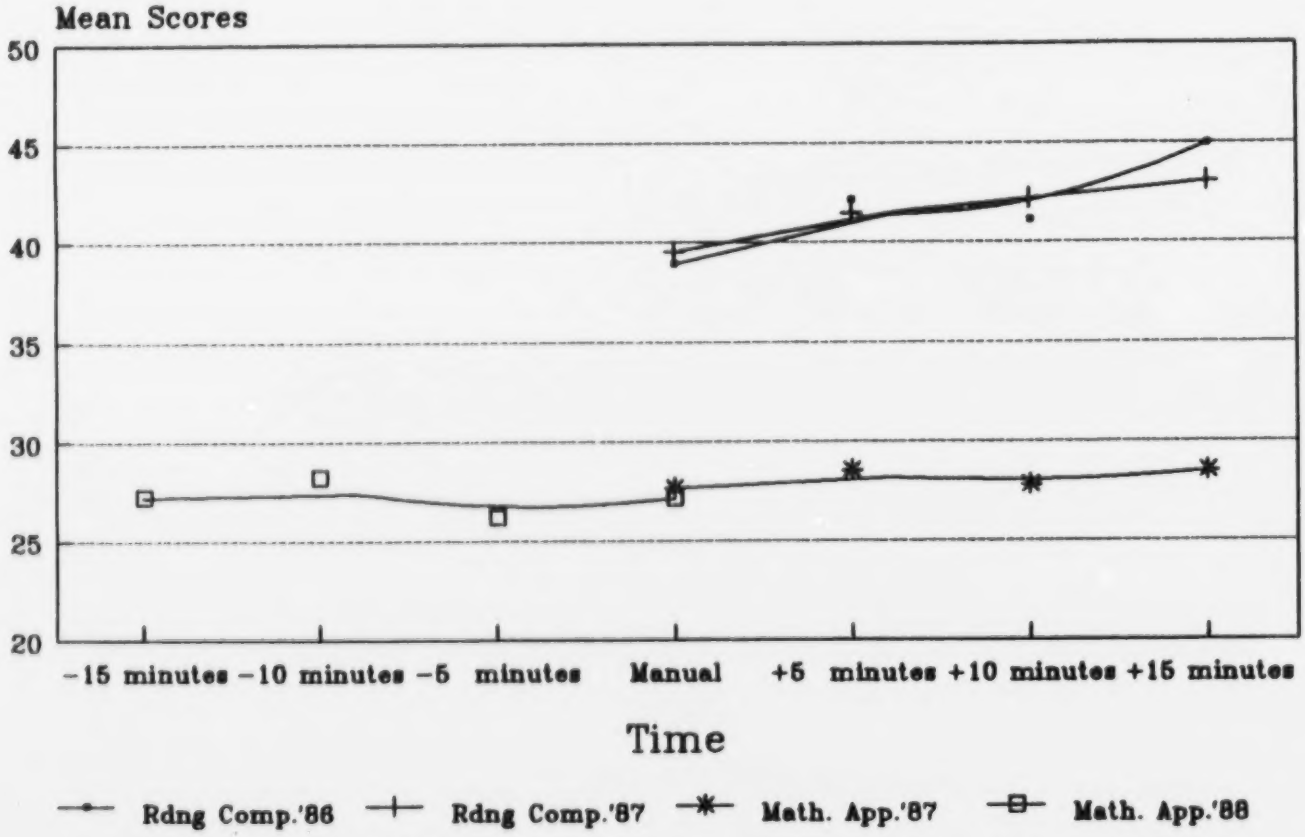
Subtest	Treatment	Raw Score			Scaled Score			Grade Equivalent		
		I	II	III	I	II	III	I	II	III
Reading Comp.	1	39	40	---	632	635	---	5.0	5.2	---
	2	43	41		645	638	---	5.9	5.4	---
	3	41	42		638	642	---	5.4	5.7	---
	4	45	43		652	645	---	6.4	5.9	---
Math Applications	1	---	28	27	---	630	625	---	5.4	5.1
	2	---	29	26	---	635	621	---	5.6	4.9
	3	---	28	28	---	630	630	---	5.4	5.4
	4	---	29	27	---	635	625	---	5.6	5.1

[1] Madden, R. , Gardner, E.F., Rudman, H.C., Karlsen, B., Merwin,J.C., Callis, R. and Collins,C. (1993). Stanford Achievement Test Series: Multi-level Norms Booklet--National. Cleveland, OH: The Psychological Corporation. P. 39, 164-165. Modified table reprinted by permission of the publisher.

NOTE: These data are only appropriate for Treatment 1 since these norms are based on the maximum times established by the authors. All other data are only illustrative of growth or lack of growth of test scores as observed in the experiments described in this study.

Effect of Time on Test Scores

Reading Comp. and Math App.



Fifth grade, 1986-1988; N=1,219

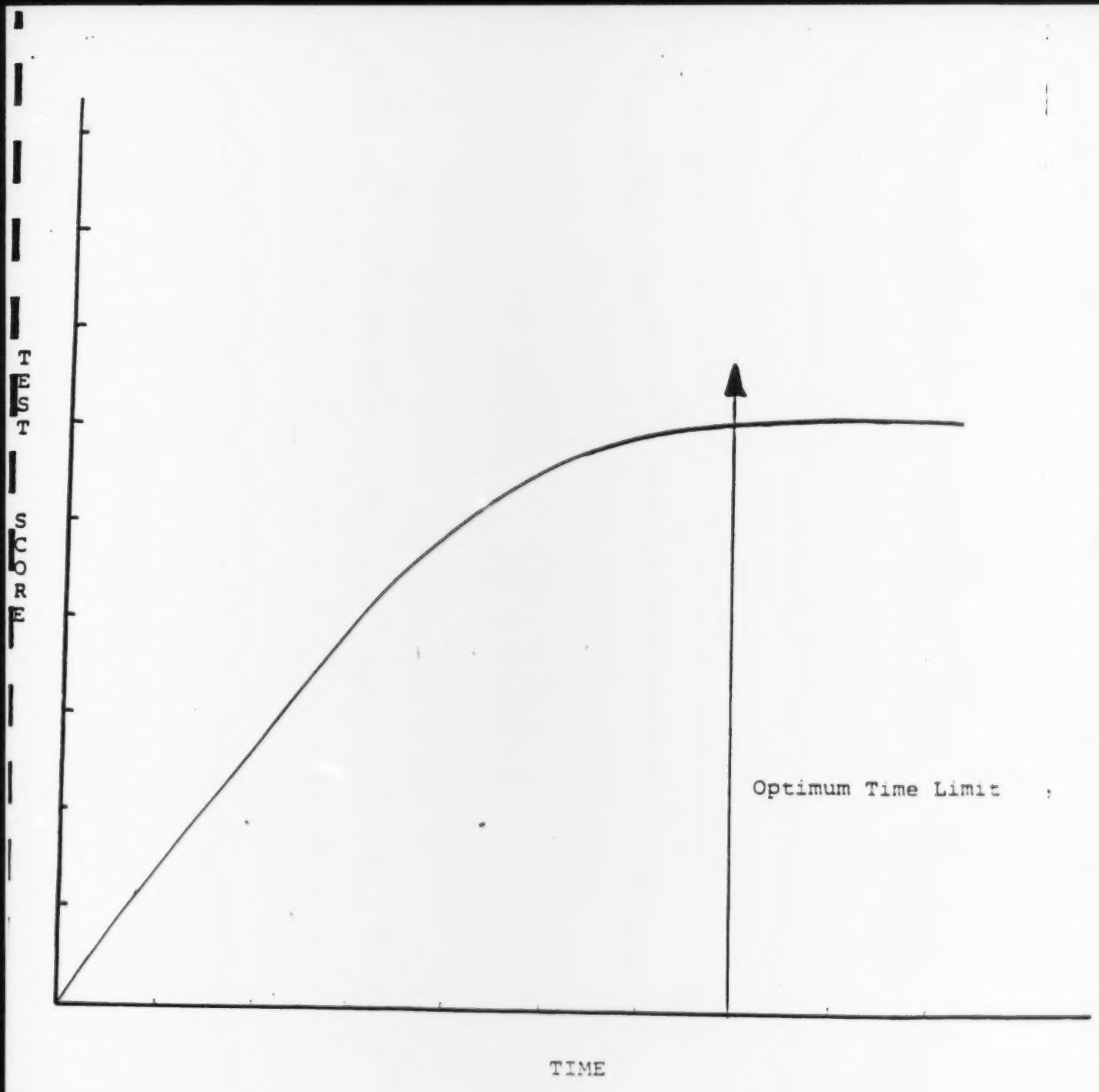


Figure 1: A Model for Determining the Optimal Time Limit of Timed Power Tests

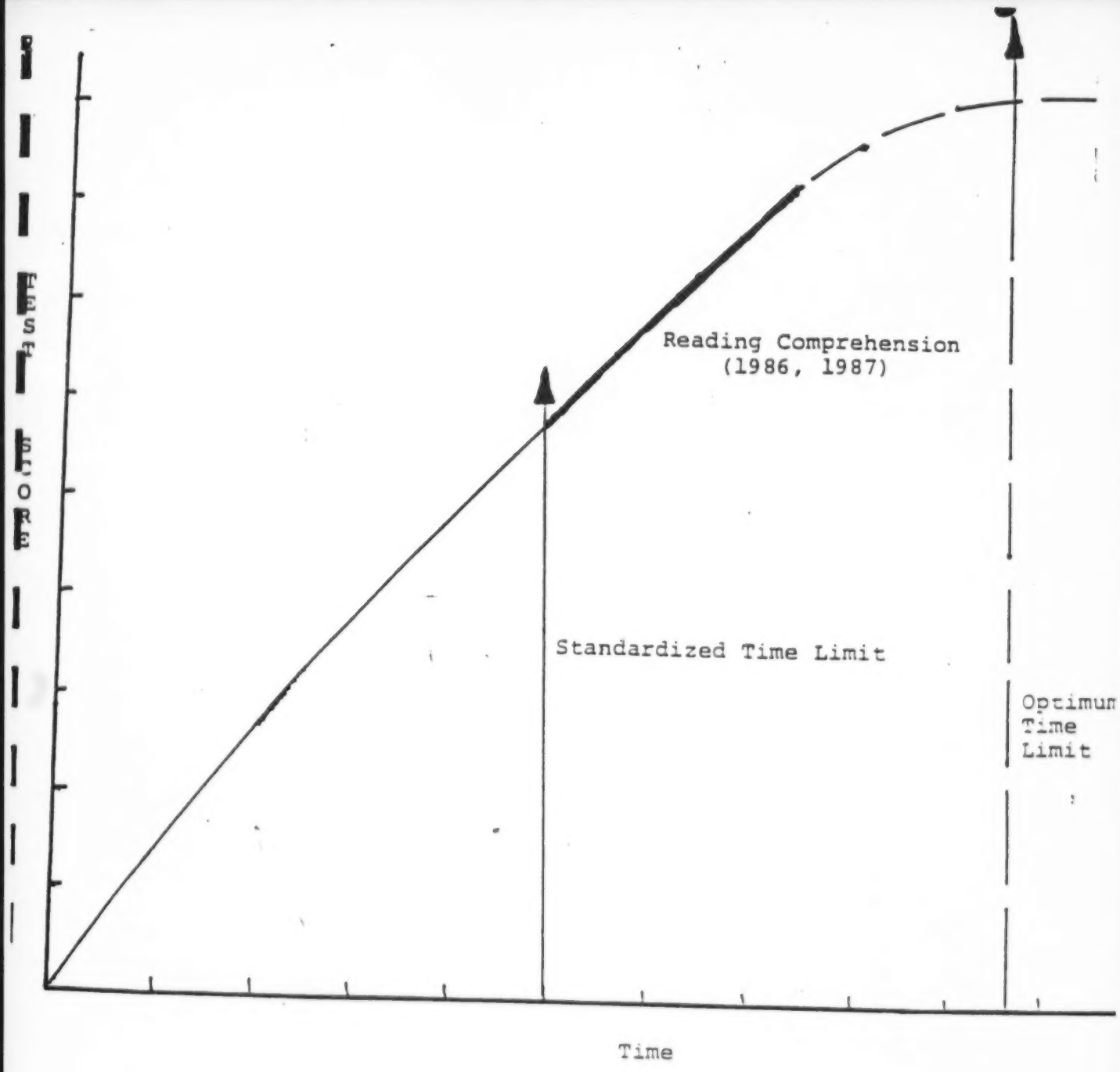


Figure 2: An Illustration of a Subtest That is Sensitive to Excess Testing Time

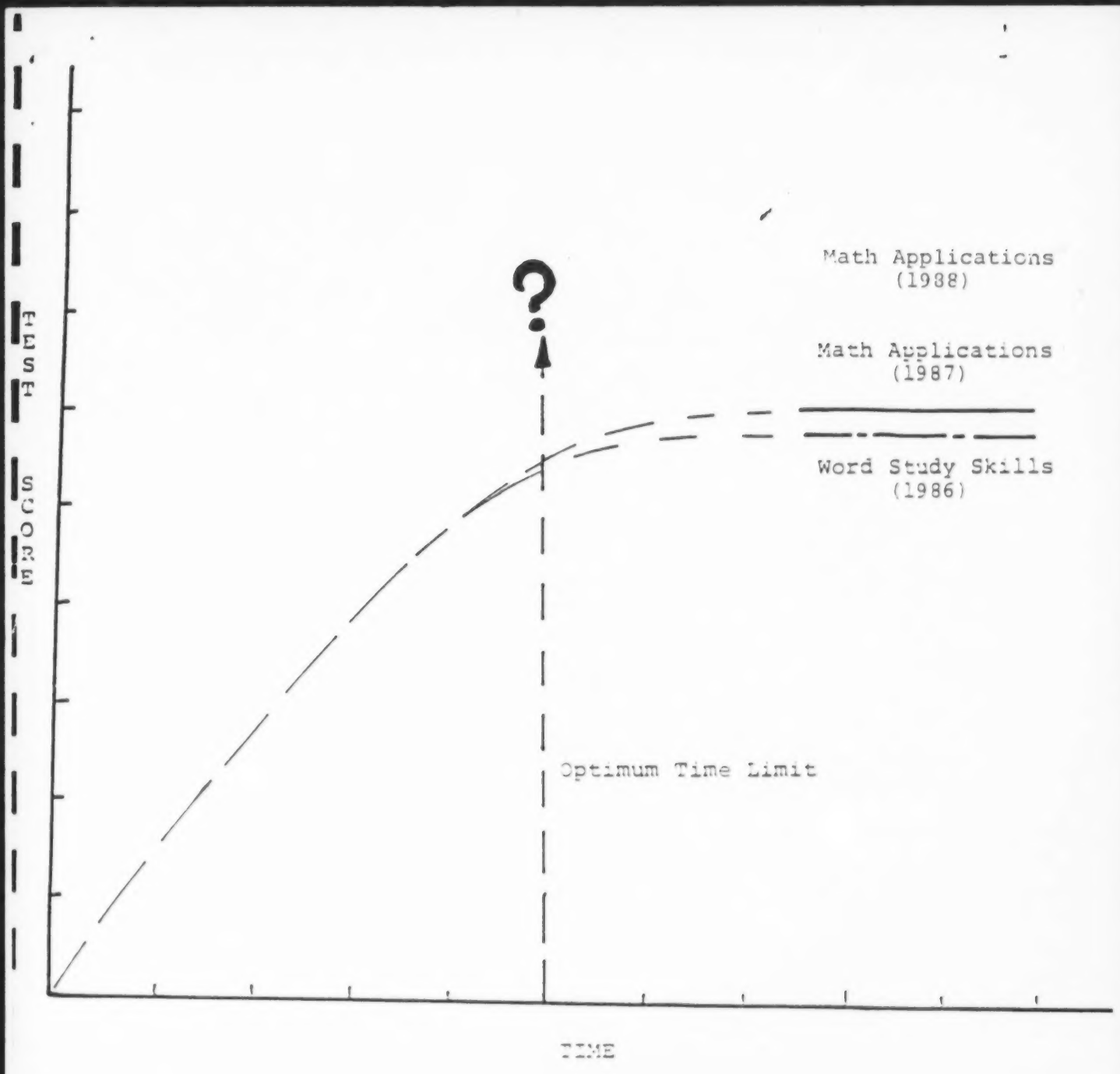


Figure 3: An Illustration ofTwo Subtests
Not Sensitive to Excess Time